DOCUMENT RESUME

ED 381 994 FL 021 916

AUTHOR Weigle, Sara Cushing
TITLE Using FACETS To Model Rater Training Effects. Draft.
PUB DATE Feb 94
NOTE 40p.; Paper presented at the Language Testing Research Colloquium (Washington, DC, 1994).
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS English (Second Language); *Essay Tests; *Evaluation Criteria; *Evaluators; *Interrater Reliability; Language Tests; Rating Scales; Second Language Learning; Testing
IDENTIFIERS *FACETS Computer Program; *University of California Los Angeles

ABSTRACT

This paper describes a study on rater training that involved the analysis of ratings given to English-as-a-Second-Language (ESL) compositions by 8 inexperienced and 8 experienced raters both before and after rater training, using FACETS (Linacre, 1990, 1993), which provides measures of rater severity and consistency. The testing text was a 50-minute composition essay, with 2 prompts, from the ESL Placement Examination (ESLPE) at the University of California, Los Angeles. Compositions were rated using the ESLPE Rating Scale on content, rhetorical control, and language. Each essay was read by two raters, primarily ESL faculty and teaching assistants, and the scores averaged. All raters attended mandatory composition rater training. FACETS provided a 4-faceted model based on estimates of examinee ability, rater harshness, scale difficulty, and prompt difficulty. Pre-training, all raters as a group differed quite significantly from one another in terms of severity. Post-training, a clear distinction between new and old raters is no longer visible. Findings indicate that rater severity evened out somewhat after training across the group. However, the spread of rater severities after training is still quite significant. Rater consistency improved, and new rater extremism was reduced. Results confirm that rater training cannot make raters into duplicates of one another, but it can make them more self-consistent. Appendixes include the ESLPE rating guidelines and sample ESLPE. (Contains 25 references.) (NAV)

Discussion Draft: Comments are welcome

Using FACETS to Model Rater Training Effects

LTRC 1994, Washington, D.C.

Sara Cushing Weigle
Department of TESL & Applied Linguistics
University of California, Los Angeles

February 1994

Introduction

Very few, if any, testing specialists would argue with the proposition that rater training is essential in achieving reliable ratings of essay examinations. However, in both the writing assessment and measurement literature some controversy exists as to the purpose and efficacy of rater training. In the writing assessment literature, concerns have been voiced about the validity of holistic essay examination scores because of the artificiality of the procedures used to reach acceptable rater reliability, including training (Charney, 1984; Gere, 1980; Barritt et al., 1986, Huot, 1990). It has been argued that an emphasis on rater consensus may force raters to ignore their own experience and expertise in judging writing, which are viewed as essential components of the interactive reading process (Barritt et al., 1986), and that inter-rater agreement may only be possible when raters are agreeing on superficial aspects of the text (Charney, 1984).

On the other hand, essay scoring and rater training procedures are presumably founded on the premise that an

essay examination is measuring a particular ability which can be defined operationally and measured accurately if raters can be trained to agree on the definition of the ability. From this point of view, it is essential for raters to put aside their own subjective experience in order to adopt the agreed-upon scoring criteria for the examination. Thus a tension exists in the writing assessment literature between these two viewpoints on the functions of rater training in writing assessment.

In the measurement literature, the function of rater training has been addressed from a somewhat different perspective, particulary with regard to the desirability and feasibility of obtaining high reliability in ratings. Linacre (1989) notes that true-score approaches to the problem of rater variation see such variation as undesirable error variance which must be eliminated or reduced as much as possible, and he outlines a number of practical and theoretical problems with achieving this goal. For instance, even if judges could be trained to give exactly the same scores to a given examinee (which has so far proven to be an impossible task), questions about the interpretability of test scores would still remain: the rating scale may not be linear, a score of "2" on one task may not mean the same as a score of "2" on another task, and so on. Although these are empirical questions, they are not routinely investigated by test users.

The many-faceted Rasch model developed by Linacre takes a different approach to the phenomenon of rater variation. In this approach, rater variation is seen as an inevitable part of the rating process and, rather than being a hindrance to measurement, is considered actually beneficial because it provides enough variability to allow probabilistic estimation of rater severity, task difficulty, and examinee ability on the same linear scale.

Proponents of the Rasch approach to measurement claim that raters cannot be trained to achieve similar levels of severity (Lunz, Wright, & Linacre, 1990). In fact, the use of Rasch measurement obviates the need to bring raters into close agreement, since estimates of examinee ability are said to be independent of the severity of the particular raters who happen to rate those particular examinees. Training (or retraining) is recommended, however, for raters who are identified as misfitting by the Rasch analysis (Lunz, Wright & Linacre, 1990; Stahl & Lunz 1991). The implication is that the function of training is not, or should not necessarily be, to force raters into agreement with each other (interrater reliability), but rather to train raters to be self-consistent (intrarater reliability).

This view of the function of training is laudable in that it allows for some variability in rater reactions to a text, which is a natural part of the reading process (Stock & Robinson, 1987). This addresses the concern raised above that holistic scoring procedures disallow personal reactions

for the sake of rater consensus and thus prevent a complete reading of the text. However, a deemphasis on interrater agreement may have implications for the construct validity of the test if it draws attention away from getting raters to agree on a definition of the ability being measured by the test. In other words, if internal consistency is considered the most important benefit of training and differences in rater severity are compensated for mathematically, a thorough understanding of the intended grading criteria may no be longer a central aspect of the training process. Raters may in fact learn to interpret the scoring rubric in idiosyncratic ways, each of which may be consistent in itself but which may not have anything to do with the construct of writing ability as defined by the test writers.

Despite discussions in both the composition literature and the measurement literature about the functions of rater training, little is known about what actually occurs during rater training and how it affects the raters themselves. Can training induce raters to be more or less severe in their judgments, as suggested by Freedman (1981), or is rater severity a stable characteristic which differs from rater to rater, as Lunz, Stahl, and Wright (1991) claim? Can training at best be seen as a way to make raters more internally consistent, so that their innate differences of severity can be compensated for mathematically? Equally importantly, to what extent does rater training function to

bring raters into agreement about the definition of the ability which the test is intended to measure?

Although a few studies have looked at differences between trained and untrained raters in writing assessment (Huot, 1988; Cumming 1990; Shohamy, Gordon, & Kraemer, 1992) and other contexts (e.g. Myford 1991a, 1991b), few if any studies have used a pre- and post- training design, as Hamp-Lyons (1990) notes. In addition, there have been very few studies of the influence of training in second language writing assessment. Hamp-Lyons (1990) deplores this gap, stating that

> we need to understand a great deal more than we do at present about the way training should be conducted, what aims it should have, the interaction between the values embodied in training and the values implicit in the task, and the criteria (defined and implied.) (p. 81)

Clearly there is a need for a better understanding of the process of rater training in assessment in general, and for first and second language writing assessment in particular. This study attempts in a small way to address this need.

Context of the Study

The study described in this paper involves the analysis of ratings given to English as a Second Language (ESL) compositions by both inexperienced and inexperienced raters both before and after rater training. The analysis was done using FACETS (Linacre, 1990, 1993), which provides measures

of rater severity and consistency, to be described later in this paper.

The testing context for the study is the composition subtest of the English as a Second Language Placement Examination (ESLPE) given quarterly at the University of California, Los Angeles (UCLA). The ESLPE composition subtest consists of a 50-minute essay on the student's choice between two prompts. One prompt requires students to interpret a graph and make predictions based on the information from the graph (the GRAPH INTERPRETATION prompt), while the other prompt requires students to make and defend a choice based on information contained in a chart or table (the CHOICE JUSTIFICATION prompt). At the time that this study was carried out, three CHOICE JUSTIFICATION prompts and three GRAPH INTERPRETATION prompts were used in rotation on different forms of the ESLPE, but each form contained one of each kind of prompt.

The compositions are rated using the ESLPE Rating Scale, which consists of three 10-point subscales (Content, Rhetorical Control, and Language). Each scale is divided into five two-point bands with descriptors for each band. (See Appendix A.) The total score is derived by doubling the language score and summing all three subtest scores. Each essay is read by two raters, and the two scores are averaged. In cases of extreme score differences (five or more points), the essay is given to a third reader, and the two scores

which are closest to each other are used in determining the final score.

The compositions are rated primarily by ESL faculty and teaching assistants (TAs). All TAs are expected to participate in the composition rating during each quarter that they are employed; however, they are allowed to arrange for a substitute if they cannot or do not wish to participate. Thus, while most composition raters are ESL teachers at UCLA, occasionally raters are hired with little or no experience or interest in ESL teaching.

All raters attend mandatory composition rater training, or "norming" sessions, which are led by the composition supervisor. The most extensive norming sessions, each lasting about two hours, take place during the Fall quarter, when many new raters must be hired.

A typical norming session at UCLA is conducted as follows. Before the norming session, raters are given norming packets, which include representative samples of compositions from previous administrations of the ESLPE. These compositions have been scored by ESL section faculty, and a consensus score for each subscale is written on each composition.

The compositions in the norming packet span the different bands of the ESLPE rating scale and include sample compositions from each writing prompt that the raters will encounter in the live reading. However, not every point on each subscale is represented in the norming packets; for

example, in the Fall 1992 norming packets there were no examples of essays that received a score of '10' on any subscale.

Before the norming session, raters are asked to score the compositions and compare their scores to the "official" scores. During the norming session itself, the supervisor asks for questions about the compositions and the official scores that have been given. In cases where raters have given a different score, they are encouraged to explain why they have given a particular score, and to reach an understanding of the rationale behind the official score. Although a complete description of the norming sessions is beyond the scope of this paper, it is important to note that a certain amount of deviation from the official score is accepted, since, as mentioned above, the reported score is an average of two ratings.

Although the norming session itself is the principal means of rater training, a certain amount of socialization of raters occurs during the live rating as well, as all rating is done in a group setting. For example, after scoring an essay the second rater on each essay can see the ratings given by the first rater and receive feedback on his or her rating in this way. Also, at times during the live rating the supervisor may choose to speak to individual raters if she senses that their rating is aberrant in some ways. For the purposes of this paper, then, the phrase "training process" refers both to the formal training received in the

norming session, and the informal training received through the live rating.

Research questions

The research questions for the study were as follows:

(1) To what extent does rater training affect the severity of individual raters?

(2) To what extent does rater training affect the spread of rater severities (i.e. make raters more like each other in terms of severity)?

(3) To what extent does rater training make raters more consistent in their judgments?

(4) To what extent do experienced and inexperienced raters differ in their severity and consistency before and after rater training?

Subjects

The subjects for the study consisted of 16 people, eight from each of the following groups:

A. Raters who had never rated compositions at UCLA and thus had never been exposed to the ESLPE composition prompts, student essays, or scoring guide and procedures used in scoring the ESLPE (hereinafter: NEW). These subjects included seven new Teaching Assistants at UCLA with zero to ten years of teaching experience, all of whom had been hired to teach ESL courses during Fall quarter, 1992, and one

graduate student in Applied Linguistics who wished to learn to rate compositions. Two of these raters had had experience with composition rating, but not with the ESLPE. All eight NEW raters were female, and all were native speakers of English, with two exceptions: the non-TA was a native speaker of Navajo and another NEW rater was a native speaker of Korean, but both had learned English as young children and had native-like proficiency.

B. Experienced raters who had rated ESLPE compositions using the ESLPE rating scale before and thus were quite familiar with the rating scale, the composition prompts, and the level of student writing commonly found at UCLA (hereinafter OLD). This group consisted of returning UCLA TAs, most of whom had had at least two years of prior experience with the rating scale and from two to ten years of teaching experience. Six of these raters were male and two were female. All were native speakers of English.

There were two main reasons for choosing these two groups. First of all, the groups represent the two major backgrounds of typical ESLPE composition raters and thus constitute a sample that is representative of the population to which the study is intended to generalize. A second and more important reason for choosing these groups was to investigate any differences between inexperienced and experienced raters in terms of how they approach the task of composition rating and how they are affected by the training

process, with particular focus on the NEW raters before and after training.

Subjects were selected from the list of 1992-1993 Teaching Assistants in the ESL Section at UCLA. All potential subjects were asked about their willingness to participate in the study in exchange for monetary or in-kind compensation; all those who agreed were included in the study.

Materials

The materials for the study included the following:

ESLPE Composition: A stratified random sample of sixty compositions from the Fall 1991 administration of the ESLPE were used. These were evenly divided between two of the composition prompts (one CHOICE JUSTIFICATION and one GRAPH INTERPRETATION) that were scheduled for use on the Fall 1992 ESLPE. These prompts are found in Appendix B. The compositions were selected to represent all levels of proficiency among ESLPE examinees, based on the original scores assigned to each composition.

ESLPE Rating Scale: The ESLPE Rating Scale, as described above, was used for all composition rating.

Data Collection

Step 1 (BEFORE TRAINING)

Data collection began approximately two weeks before the norming sessions for the Fall 1992 ESLPE composition rating

took place. Several different but overlapping subsets of 15 essays from each set of essays (CHOICE and GRAPH) were created and assigned to raters, so that each rater rated 15 CHOICE essays and 15 GRAPH essays and each essay was read by six to eight raters, with the following exception. Four of the essays (two on each prompt) were read by all sixteen raters using talk-aloud procedures (Ericsson & Simon, 1984); in other words, raters spoke their thoughts into a tape recorder as they rated the essays. These ratings were collected for a qualitative analysis of rater behavior which is described elsewhere (Cushing, 1993; Weigle, forthcoming.)

The design was counterbalanced so that half of the subjects read the CHOICE JUSTIFICATION essays first, and half read the GRAPH INTERPRETATION essays first. Essays were presented in a different random order for each subject. The ESLPE Rating Scale was used for all ratings.

Step 2 (RATER NORMING)

All subjects participated in one of the two regular norming sessions for the Fall 1992 ESLPE, which lasted approximately 90 minutes each. Subjects also participated in the operational rating session of the ESLPE compositions, but the data from these rating sessions were not used for this study. The ESLPE was administered five times over the period of ten days, and composition rating took place the day of the exam and on the following day. Subjects reported having

spent from six to ten hours rating compositions over the course of the ESLPE scoring sessions.

Step 3 (AFTER TRAINING)

The final data collection took place within one to three weeks following the operational ESLPE rating. Subjects rated different but overlapping sets of 16 essays on each of the two prompts, using the ESLPE Rating Scale. Most of these essays were essays that the raters had not rated in the PRE data collection. Six of the essays, including the four that all subjects had previously rated while speaking their thoughts aloud, were rated using talk-aloud procedures. Subjects were asked to read the essays within a week. Most subjects complied with this request, although a few subjects took up to three weeks to finish the rating.

Data Analysis

Rater behavior before and after training was modelled using the IRT program FACETS, which provides estimates of examinee ability, rater harshness, scale difficulty, and prompt difficulty on a common log-linear metric or logit scale. The mathematical model in a four-faceted model with facets of examinee, prompt, rater, and scale can be expressed as follows:

$$\log (P_{njikx}/P_{nrjik-1}) = Bn - Cj - Di - Ek - Fx$$

where P_{nrpsx} = the probability of examinee n being awarded a score of category x on prompt i on scale k by rater j
$P_{nrpsx-1}$ = probability of examinee n being awarded a score of category x-1 on prompt i on scale k by rater j
Bn = ability of examinee n
Cj = severity of rater j
Di = difficulty of prompt i
Ek = difficulty of scale k
Fx = threshold difficulty of being rated in category x relative to category x-1.

In addition to providing logit estimates of the ability, severity, or difficulty of each element of each facet, FACETS also provides statistics indicating the relative spread of these estimates within each facet. In other words, the analysis provides information about the significance of any differences that may exist among elements of a facet; for example, differences in severity among raters or ability among examinees.

Another important feature of the FACETS analysis is that it provides fit statistics for each element, which provide an indication of the degree to which each element is behaving in a manner that is predicted by the model. In the case of raters, the fit statistics are indicators of rater consistency. Thus a detailed picture of the behavior of each rater in terms of both severity and consistency can be formed.

The FACETS analysis for the PRE and POST data were performed separately. In each analysis, four facets were

used: PERSONS (i.e.,examinees, n=60), RATERS (eight NEW, eight OLD), PROMPTS (CHOICE and GRAPH), and SCALES (Content, Rhetorical Control, and Language). Logit values for the two prompts were anchored at pre-set values of .07 and -.07, respectively, based on a preliminary analysis which showed that the CHOICE JUSTIFICATION prompt was somewhat more difficult than the GRAPH INTERPRETATION prompt. (See Weigle, forthcoming, for further details.) This was necessary in order to provide enough connection in the dataset for the analysis to run, since each examinee wrote on one prompt only.

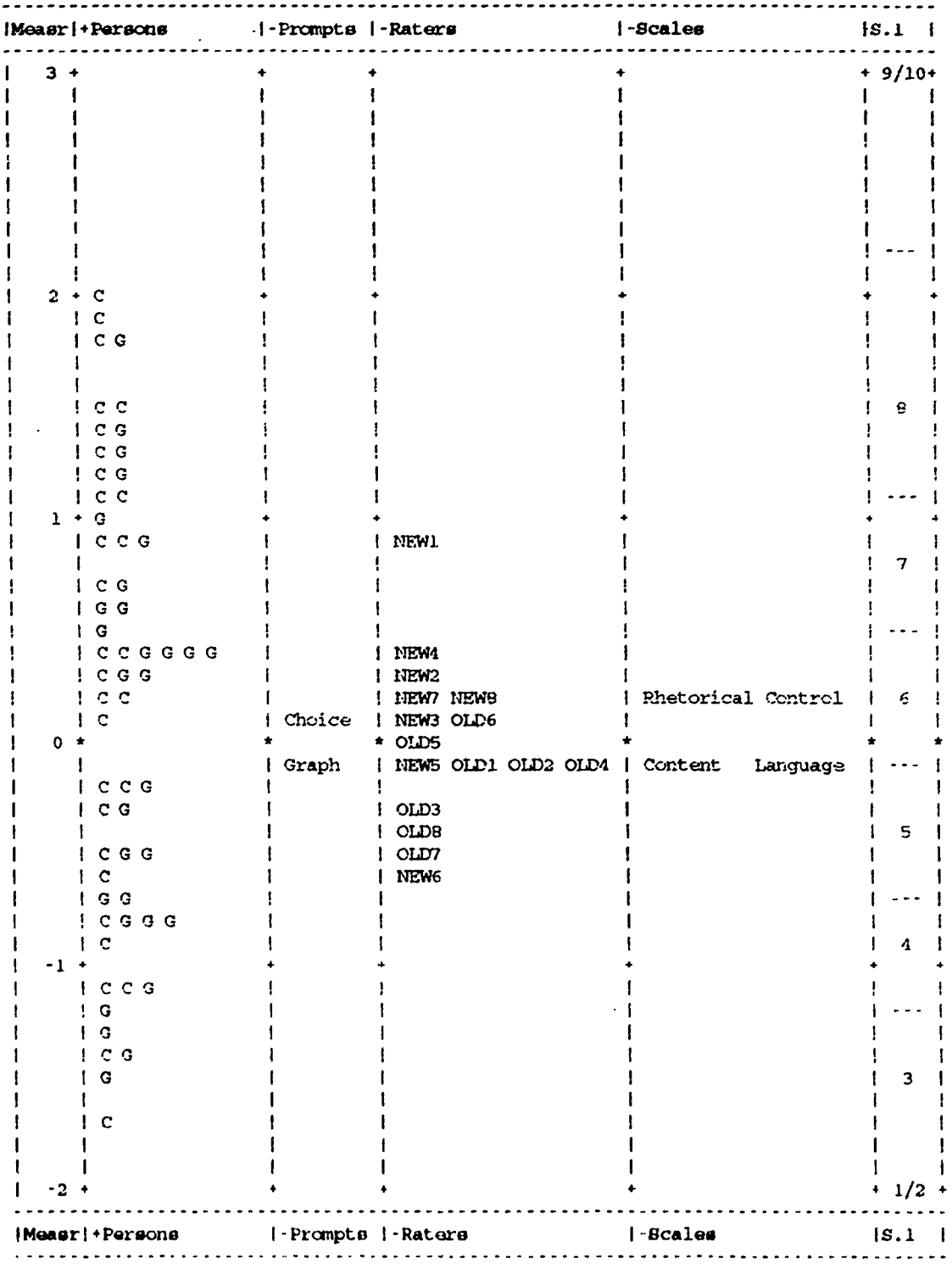
The POST analysis was performed first, as it was assumed that the ratings given to these examinees by recently trained raters would be more accurate than those given by untrained or inexperienced raters. A preliminary investigation of the data revealed that the two most extreme scores (1 and 10) were rarely given on any subscale; for this reason, and to simplify the analysis, scores of 1 and 2, and scores of 9 and 10 were combined, to make an eight-point scale. For the sake of consistency, this procedure was followed in the PRE analysis as well. For all analyses, the Rating Scale model (Andrich, 1978a, 1978b) was used, which assumes that the steps of the rating scale are equivalent across all elements of a given facet.

Results

As an overall introduction to the PRE FACETS results, Figure 1 shows graphically the measures for Persons, Prompts, Raters, and Scales for the PRE data. The figure is to be interpreted as follows. The scale along the left of the figure represents the logit scale, which is the same for all facets. Each person is represented by the letter C (for CHOICE JUSTIFICATION) or G (for GRAPH INTERPRETATION) depending upon which prompt that person wrote on. Persons are ordered with the most able examinees at the top, and the least able at the bottom. The other facets are ordered so that the most difficult element of each facet is towards the top, and the least difficult towards the bottom. In terms of Raters, the most severe rater (Rater NEW1) is the uppermost rater in the figure. Similarly, the most difficult Scale (Rhetorical Control) is uppermost in the figure. Finally, the most likely scale score for each ability level is shown in the farthest right column. The figure thus shows pictorially the differences across the different elements of each facet.

Figure 1 about here

Figure 1. PRE: All Facet Summary.



As the figure indicates, examinee estimates range from a high of about 2 logits to a low of close to -2 logits. The facet of Prompt (anchored) shows that the CHOICE JUSTIFICATION prompt is slightly more difficult than the GRAPH INTERPRETATION prompt. The scale for Rhetorical Control is more difficult than either Content or Language Looking at the column for Raters, we see that Rater NEW1 is by far the most severe rater. Of the raters clustered around the mean (0), the NEW raters tend to be more severe than the OLD raters. Finally, four raters are more lenient than the rest, with Rater NEW6 the most lenient of all. Thus the figure clearly shows that the raters, before training, are not at the same level of severity. This is particularly true of the NEW raters, which is as we might expect.

RATER SEVERITY

A more detailed analysis of rater behavior is found in Table 1, the Raters Measurement Report for the PRE data. Raters are presented in descending order of severity; in other words, Rater NEW1 is the most severe and Rater NEW6 is the least severe, as was seen in Figure 1. The most striking thing to be noticed in this table is the fact that five out of the eight NEW raters are more severe than the most severe of the OLD raters, and a seventh (Rater NEW6) is the most lenient. Only Raters NEW3 and NEW5 fall within the severity range of the OLD raters. What this suggests is that the NEW raters tend to be more extreme in their ratings than the OLD

raters, and the majority of the NEW raters are applying stricter standards overall to the essays than are the OLD raters.

Table 1. PRE: Raters Measurement Report

Raters	Measure Logit	Model Error	Infit MnSq	Infit Std
NEW1	0.94	0.09	1.9	4
NEW4	0.36	0.08	0.9	-1
NEW2	0.26	0.08	0.6	-3
NEW7	0.24	0.08	2.1	5
NEW8	0.17	0.08	0.8	-1
OLD6	0.11	0.08	0.8	-1
NEW3	0.10	0.08	1.5	2
OLD5	-0.00	0.08	0.9	0
OLD1	-0.07	0.08	0.6	-2
OLD2	-0.12	0.08	1.1	0
NEW5	-0.13	0.08	0.8	-1
OLD4	-0.13	0.09	0.6	-3
OLD3	-0.30	0.08	0.6	-3
OLD8	-0.40	0.09	0.9	0
OLD7	-0.46	0.09	0.8	-1
NEW6	-0.57	0.09	0.7	-2
Mean	-0.00	0.08	1.0	-0.6
S.D.	0.36	0.00	0.4	2.7

RMSE 0.09 Adj S.D. 0.35 Separation 4.08 Reliability 0.94
Fixed (all same) chi-square: 266.89 d.f.: 15 significance: .00
Random (normal) chi-square: 14.95 d.f.: 14 significance: .38

The FACETS analysis provides a number of indications of the magnitude of the differences among elements of a facet: in this case, in severity among raters. These are the Separation Index, the Reliability, and the Fixed (all same) chi square. The separation index is the ratio of the corrected standard deviation of element measures (in this case, Raters) to the root mean-square estimation error. If the Raters were equally severe, the standard deviation of the Rater difficulty estimates should be equal to or smaller than

the mean estimation error of the entire data set. However, the Rater Separation Index is 4.08, indicating that the variance among raters is about four times the error of estimates.

The Reliability statistic provided by the FACETS analysis indicates the degree to which the analysis reliably distinguishes between different levels of difficulty or severity among the elements of the facet (in this case, the different raters). For raters (in fact, for all facets other than examinees) a low reliability is desirable, since ideally the different raters would be equally severe. In this case, however, the reliability is .94, indicating that the analysis is fairly reliably separating raters into different levels of severity. Finally, the Fixed chi-square tests the null hypothesis that all of the elements of the facet are equal. The chi-square of 266.89 with 15 d.f. is significant at p=.00, indicating that the null hypothesis must be rejected; in other words, the raters are not equally severe.

RATER CONSISTENCY

The FACETS analysis also provides two measures of fit, or consistency: the infit and the outfit. The infit is the weighted mean-squared residual which is sensitive to unexpected responses near the point where decisions are being made, while the outfit is the unweighted mean-squared residual and is sensitive to extreme scores. For the purposes of this study, only the infit statistics will be

reported. Although there are no hard and fast rules for determining what degree of fit is acceptable, some researchers (e.g., Lunz & Stahl, 1990) have found the lower and upper limits of .5 and 1.5, respectively, for mean squares to be useful for practical purposes. Fit statistics 1.5 or greater indicate too much unpredictability in raters' scores, while fit statistics of .5 or less indicate overfit, or not enough variation in scores.

In addition to the mean squares, FACETS provides standardized infit statistics, which have an expected mean of 0 and standard deviation of 1. These statistics are useful for comparing the elements of a facet with each other, as they show the degree of variability in individual raters' ratings relative to the amount of variability in the entire data set. Standardized fit statistics greater than 2 or less than -2 are generally signs of misfit.

Applying these standards to Table 4.3, we can see that three out of the eight NEW raters had very high infit statistics (≥ 1.5): Raters NEW1, NEW3, and NEW7. These statistics indicate that these raters' ordering of examinees was not consistent with the estimated ability measures of the examinees, and that the scores that they gave were highly unpredictable.

OLD vs. NEW RATERS

Table 2 shows the Raters Measurement Report for the two groups of raters: NEW and OLD. A number of differences

between groups can be seen from the table. First of all, the NEW raters have a mean severity of .17 logits, while the OLD raters as a group are less severe, with a mean of -.17 logit.

Table 2 Raters Measurement Report

Raters	Measure Logit	Model Error	Infit MnSq	Infit Std	Raters	Measure Logit	Model Error	Infit MnSq	Infit Std
NEW1	0.94	0.09	1.9	4	OLD1	-0.07	0.08	0.6	-2
NEW2	0.26	0.08	0.6	-3	OLD2	-0.12	0.08	1.1	0
NEW3	0.10	0.08	1.5	2	OLD3	-0.30	0.08	0.6	-3
NEW4	0.36	0.08	0.9	-1	OLD4	-0.13	0.09	0.6	-3
NEW5	-0.13	0.08	0.8	-1	OLD5	-0.00	0.08	0.9	0
NEW6	-0.57	0.09	0.7	-2	OLD6	0.11	0.08	0.8	-1
NEW7	0.24	0.08	2.1	5	OLD7	-0.46	0.09	0.8	-1
NEW8	0.17	0.08	0.8	-1	OLD8	-0.40	0.09	0.9	0
MEAN	0.17	0.08	1.2	0.5	MEAN	-0.17	0.09	0.8	-1.6
S.D.	0.40	0.00	0.5	3.2	S.D.	0.19	0.00	0.2	1.4

NEW RATERS:
RMSE 0.08 Adj S.D. 0.39 Separation 4.62 Reliability 0.96
Fixed (all same) chi-square: 165.54 d.f.: 7 significance: .00
Random (normal) chi-square: 6.98 d.f.: 6 significance: .32

OLD RATERS:
RMSE 0.09 Adj S.D. 0.17 Separation 1.94 Reliability 0.79
Fixed (all same) chi-square: 37.95 d.f.: 7 significance: .00
Random (normal) chi-square: 6.99 d.f.: 6 significance: .32

Because of the small sample size, parametric statistical tests are not appropriate to test the significance of this difference. A Mann-Whitney U test, a non-parametric test based on the rank order of rater severities, was therefore performed in order to test the null hypothesis that the two groups of raters are equal in rater severity. The results of this test are found in Table 3.

Table 3. Mann-Whitney U test - PRE OLD vs. NEW

Rater type	n	S Rank	Mean Rank	Z
NEW	8	86.5	10.81	-1.943
OLD	8	49.5	6.19	

Z = -1.943, p = .052

As the table indicates, the difference between the two groups of raters does not quite reach statistical significance at a = .05, although it is quite close. This result is due primarily to the presence of Rater NEW6, the least severe rater, who is unlike the other NEW raters in that she is quite lenient in her ratings. However, the overall trend towards severity among the NEW raters remains intriguing.

The NEW raters as a group also vary much more in their severity, with a standard deviation of .40 logits, compared with .19 for the OLD raters. This variability is also reflected in the separation indices for the two groups: 4.62 for the NEW raters, and 1.94 for the OLD raters. However, it should be noted that the OLD raters also differ significantly in their severity, indicating that despite their greater experience with the ESLPE rating scale, they still do not rate compositions in the same way.

In terms of consistency, a number of group differences can be seen as well. The standardized infit has an expected

mean of 0 and S.D. of 1; note that for the NEW raters these figures are .5 and 3.2, respectively, while for the OLD raters they are -1.6 and 1.4. Thus the NEW raters are much less consistent, and more unlike each other in terms of consistency than would be expected, while the OLD raters are more consistent than expected.

To summarize the results of the PRE analysis, then, we have seen that all raters as a group differ quite significantly from one another in terms of severity. Three raters, all of them NEW raters, are highly inconsistent in their ratings. Most of the NEW raters are more severe than the OLD raters, and the NEW raters as a group are less consistent than the OLD raters. These results, particularly in terms of rater consistency, are not surprising, and are consonant with previous research that shows that untrained raters tend to be less reliable than trained raters (e.g., Diederich, French, & Carlton, 1961; Shohamy, Gordon, & Kramer, 1992). More intriguing is the finding that so many of the NEW raters were more severe than the OLD raters, as some previous research has shown untrained raters to be less severe overall than trained raters (Ruth & Murphy, 1988).

POST

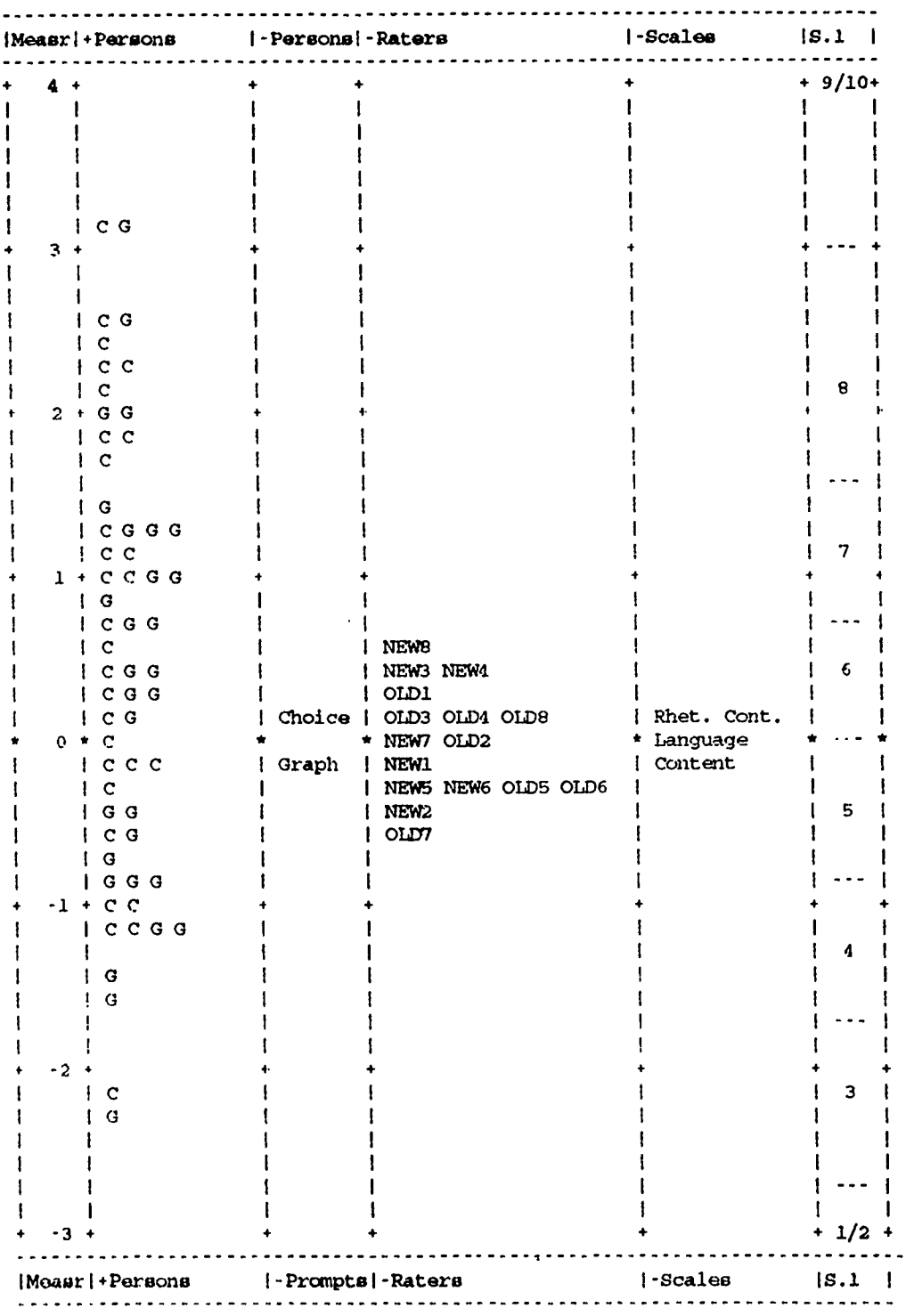
Figure 2 summarizes the FACETS results for all facets using the POST data. As before, the Persons are arranged from the most able at the top to the least able at the bottom, while the Raters are arranged from most to least

severe. A few differences from the PRE data can be seen in this figure. First of all, the logit scale for examinees has expanded, with ability estimates ranging from lower than -2 to more than 3 logits. This suggests that the raters as a group are making finer distinctions among ability levels after training than they had done before training. In addition, the Raters are clustered together around the mean more than they had been in the PRE data, suggesting that there is not as much variation among raters as there had been before training. Finally, the Scales appear to be closer together; Rhetorical Control is not as distinct from the other two scales as it had been.

Figure 2 about here

Table 4 shows the measurement report for raters using the POST data. The POST data show a quite different picture for the raters from that from the PRE data. In terms of rater severity, as can be seen most clearly in Figure 2, there is no longer a preponderance of NEW raters at the extreme ends of the severity range; a clear distinction between OLD and NEW raters can no longer be made. Furthermore, the spread of rater severities is somewhat decreased, with a separation index is 3.10, as opposed to 4.08 in the PRE data. However, note that this is still quite a significant spread, with a reliability of .91 and a highly significant chi-square. For these raters, then, the process

Figure 2. POST: All Facet Vertical Summary.



of training was effective in reducing some of the differences in rater severity, but major differences among raters remain despite training.

Table 4 POST: Raters Measurement Report

Raters	Measure Logit	Model Error	Infit MnSq	Infit Std
NEW8	0.55	0.09	1.0	0
NEW3	0.43	0.09	1.2	1
NEW4	0.39	0.10	1.6	3
OLD1	0.24	0.10	0.9	0
OLD8	0.21	0.10	1.2	1
OLD3	0.20	0.09	0.6	-3
OLD4	0.15	0.09	0.8	-1
NEW7	0.07	0.10	1.1	0
OLD2	-0.03	0.10	1.4	2
NEW1	-0.15	0.10	1.1	0
NEW5	-0.25	0.10	0.9	0
NEW6	-0.26	0.10	0.9	0
OLD6	-0.27	0.10	0.8	-1
OLD5	-0.29	0.10	0.9	0
NEW2	-0.45	0.10	0.6	-2
OLD7	-0.54	0.10	0.9	0
Mean	-0.00	0.10	1.0	-0.2
S.D.	0.32	0.00	0.3	1.9

RMSE 0.10 Adj S.D. 0.30 Separation 3.10 Reliability 0.91
Fixed (all same) chi-square: 169.84 d.f.: 15 significance: .00
Random (normal) chi-square: 15.00 d.f.: 14 significance: .38

RATER CONSISTENCY

While the overall differences in rater severity have improved only slightly, the fit statistics show quite a bit of improvement in the POST data, although not all raters fall within the limits of acceptable fit. The three NEW raters identified as misfitting in the PRE analysis (Rater NEW1, Rater NEW3, and Rater NEW7) all show improved fit statistics; Rater NEW4, on the other hand, now appears inconsistent.

However, the mean and standard deviation of Standardized Infits (-.2, 1.9) are closer to their expected value than they were in the PRE data. This indicates somewhat better model fit (Linacre & Wright, 1993), although perhaps not as good as one might hope for, due to the misfit of Rater NEW4. Still, it is an indication that the fit has improved overall from the PRE data.

OLD vs. NEW RATERS

Table 5 shows the Raters Measurement Report for the NEW and OLD rater groups. As the table shows, the two groups of raters are much closer than they were in the PRE data, with mean severity estimates of .04 and -.04 logits, respectively. (Recall that these figures were .17 and -.17 in the PRE data).

Table 5. POST Raters Measurement Report: OLD vs. NEW Raters

Rater	Measure	Model Error	Infit MnSq	Infit Std	Rater	Measure	Model Error	Infit MnSq	Infit Std
NEW1	-0.15	0.10	1.1	0	OLD1	0.24	0.10	0.9	0
NEW2	-0.45	0.10	0.6	-2	OLD2	-0.03	0.10	1.4	2
NEW3	0.43	0.09	1.2	1	OLD3	0.20	0.09	0.6	-3
NEW4	0.39	0.10	1.6	3	OLD4	0.15	0.09	0.8	-1
NEW5	-0.25	0.10	0.9	0	OLD5	-0.29	0.10	0.9	0
NEW6	-0.26	0.10	0.9	0	OLD6	-0.27	0.10	0.8	-1
NEW7	0.07	0.10	1.1	0	OLD7	-0.54	0.10	0.9	0
NEW8	0.55	0.09	1.0	0	OLD8	0.21	0.10	1.2	1
Mean	0.04	0.10	1.1	0.3	Mean	-0.04	0.10	0.9	-0.8
S.D.	0.35	0.00	0.3	1.8	S.D.	0.27	0.00	0.2	1.7

NEW RATERS:
RMSE 0.10 Adj S.D. 0.34 Separation 3.48 Reliability 0.92
Fixed (all same) chi-square: 105.39 d.f.: 7 significance: .00
Random (normal) chi-square: 7.00 d.f.: 6 significance: .32

OLD RATERS:
RMSE 0.10 Adj S.D. 0.25 Separation 2.61 Reliability 0.87
Fixed (all same) chi-square: 61.35 d.f.: 7 significance: .00
Random (normal) chi-square: 6.98 d.f.: 6 significance: .32

A Mann-Whitney U test was again performed in order to determine whether this difference between the two groups is significant. Results of this test are found in Table 6. As the table indicates, the two groups are not significantly different ($z = -.63$, $p = .529$). Thus we can see that training has reduced the differences in severity between the two groups of raters.

Table 6. Mann-Whitney U test - POST OLD vs. NEW

Rater type	n	S Rank	Mean Rank	Z
NEW	8	74	9.25	-.63
OLD	8	62	7.75	

Z = -.63, p = .529

In addition, the two groups do not differ from each other as much in terms of variability as they did in the PRE data: the s.d. for NEW raters is .35 logits, while the s.d. for OLD raters is .27. This result is due not only to a decrease in variability in the NEW raters, as expected, but, interestingly, to an increase in variability among the OLD raters. Again, this is reflected in the separation indices as well, which are 3.48 for NEW raters and 2.61 for OLD raters. While the NEW raters are more like each other POST than PRE, the opposite is the case for the OLD raters. However, this may well represent random fluctuation in severity estimates due to the small sample size (Bachman, personal communication).

Discussion

In summary, the anaysis reveals a number of changes in rater behavior from PRE to POST. First of all, the spread of rater severity estimates is somewhat reduced, indicating that raters are somewhat more like each other after training than

before. This is especially true for the NEW raters. However, it must be pointed out that the spread of rater severities after training is still quite significant, indicating that despite training raters were still markedly different from one another in their severity. The training seems to have brought the extreme scorers within a more tolerable range of severity, but beyond that it has not eliminated differences in rater severity. This is consistent with Stahl and Lunz' (1990) assertion that rater training cannot eliminate differences among raters in terms of their severity.

On the other hand, the fit statistics from PRE to POST have improved, and in particular, the three raters who were highly inconsistent before training were quite consistent afterwards. The fact that not all raters were consistent after training is some cause for concern; it may be that certain raters cannot be trained out of inconsistent rating patterns and perhaps should not be used as raters. Overall, however, rater consistency improved from PRE to POST, which is again consonant with Stahl and Lunz' assertions about the value of training.

In terms of group differences, the analysis shows that before training, the NEW raters tended to be more extreme in their severities than the OLD raters, and for the most part they were more severe as well. Although a full discussion of this phenomenon is beyond the scope of this paper, a qualitative analysis of the NEW raters' talk-aloud protocols

revealed that they tended to apply the scoring rubric more rigidly than did the OLD raters, which may have led to stricter ratings (Weigle, forthcoming). Following training, no clear group distinctions could be made between OLD and NEW raters, as individual severity estimates fluctuated from PRE to POST for both groups of raters. This suggests that the training had the effect of reducing the extremism of the NEW raters, and in particular the tendency of many of the NEW raters to judge more severely than the OLD raters.

Conclusion

Taken altogether, the results of this study partially support the claims made by Lunz, Wright, and Linacre (1990) and Stahl and Lunz (1991) that rater training cannot make raters into duplicates of each other, but it can make raters more self-consistent. This consistency will presumably make examinee measurement more accurate, as predictable variations in severity among raters can be modeled and compensated for mathematically. However, the fact that in this small sample of raters one of the eight NEW raters seems to have become less consistent after training is some cause for hesitation in interpreting the results of the study. A replication of the study with a larger sample of raters and essay scripts may show more decisive results.

Even if additional studies of this sort are carried out, many important questions regarding the functions of rater training in writing assessment remain. It may well be shown

conclusively that rater training has implications more for rater consistency than for rater severity. However, this will not resolve the dilemma posed at the beginning of this paper: that is, that the ability to compensate mathematically for differences in severity among raters may reduce emphasis on the need for agreeing on a definition of the construct being measured by the essay examination, even as it allows for the inevitable variation among raters that is inherent in the reading process. For this reason, studies such as the one presented here must be complemented by qualitative studies of rater behavior such as those of Huot (1988), Cumming (1990), Vaughan (1992), and Weigle (forthcoming), which explore the decision making processes of raters, the criteria they use in judging essays, and the match of these criteria to the intended criteria of the test designers. It is not enough to be able to assign a more accurate number to examinee performances unless we can be sure that the number represents a more accurate definition of the ability being tested.

REFERENCES

Andrich, D. (1978a). A rating formulation for ordered response categories. *Psychometrika, 43,* 561-73.

Andrich, D. (1978b). Scaling attitude items constructed and scored in the Likert tradition. *Educational and Psychological Measurement, 38,* 665-80.

Barritt, L., Stock, P., & Clarke, F. (1986). Researching practice: Evaluating assessment essays. *College Composition and Communication, 37,* 3, 315-327).

Charney, D. (1984). The validity of using holistic scoring to evaluate writing: A critical overview. *Research in the Teaching of English, 18,* 1, 65-81.

Cumming, A. (1990). Expertise in evaluating second language compositions. *Language Testing, 7,* 31-51.

Cushing, S. (1993). Effects of training on raters of ESL compositions. Paper presented at AAAL, Atlanta, GA.

Diederich, P.B., French, J.W., & Carlton, S.T. (1961). *Factors in judgments of writing ability* (Research Bulletin 61-15). Princeton, NJ: Educational Testing Service, (Eric Document Reproduction No. ED 002 172).

Ericsson, K.A., & Simon, H. (1984). *Verbal reports as data.* Cambridge, MA: MIT Press.

Freedman, S.W. (1981). Influences on evaluators of expository essays: Beyond the text. *Research in the Teaching of English, 15*(3), 245-55.

Gere, A.R. (1980). Written composition: Towards a theory of evaluation. *College English, 42,* 1, 44-58.

Hamp-Lyons, L. (1990). Second language writing: Assessment issues. In B. Kroll, (Ed.) *Second language writing: Research insights for the classroom* (pp. 69-87). Cambridge: Cambridge University Press.

Huot, B. (1988). *The validity of holistic scoring: A comparison of the talk-aloud protocols of expert and novice holistic raters.* Unpublished doctoral dissertation, Indiana University of Pennsylvania.

Huot, B. (1990). Reliability, validity, and holistic scoring: What we know and what we need to know. *College Composition and Communication, 41 ,2,* 201-13.

Linacre, J.M. (1989). *Many-faceted Rasch measurement.* Chicago, IL: MESA Press.

Linacre, J.M., & Wright, B.D. (1993). *A user's guide to FACETS (Fersion 2.6).* Chicago, IL: MESA Press.

Lunz, M.E., Stahl, J.A., & Wright, B.D. (1991). The invariance of judge severity calibrations. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Lunz, M.E., Wright, B.D. & Linacre, J.M. (1990). Measuring the impact of judge severity on examination scores. *Applied Measurement in Education, 3*(4), 331-45.

Myford, C.M. (1991a). Assessment of acting ability. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Myford, C.M. (1991b). Judging acting ability: The transition from novice to expert. Paper presented at the annual meeting of the American Educational Research Association, Chicago, IL.

Ruth, L. & Murphy, S. (1988). *Designing writing tasks for the assessment of writing.* Norwood, NJ: Ablex Publishing Corp.

Shohamy, E., Gordon, C., & Kraemer, R. (1992). The effect of raters' background and training on the reliability of direct writing tests. *Modern Language Journal.*

Stahl, J.A. & Lunz, M.E. (1991). Judge performance reports: Media and message. Paper presented at the Annual Meeting of the American Educational Research Association, San Francisco, CA.

Stock, P.L., & Robinson, J.L. (1987). Taking on testing. *English Education, 19,* 93-121.

Vaughan, C. (1992). Holistic assessment: What goes on in the rater's mind? In L. Hamp-Lyons, (Ed.) *Assessing second language writing in academic contexts* (pp. 111-26). Norwood, NJ: Ablex Publishing Corp.

Weigle, S.C. (forthcoming) Effects of training on raters of ESL compositions. Unpublished doctoral dissertation, UCLA.

REVISED VERSION, SEPTEMBER 1991

CONTENT

9-10
a. The essay fulfills the assignment well and treats the topic with sophistication. The main idea is clear.
b. Support is relevant, thorough and credible.

7-8
a. The essay addresses the assignment appropriately* and is well-developed. The main idea is clear.
b. Most of the arguments/ideas are well supported.

5-6
a. The essay addresses the topic appropriately, but may not be well-developed. OR The essay only addresses part of the topic, but develops that part sufficiently.
b. Some statements may not be supported or unrelated to main idea.

3-4
a. The essay is inappropriate to assigned topic OR the main idea is not evident.
b. The essay contains unsupported or irrelevant statements.

1-2
a. The paper lacks a clear main idea.
b. Several statements are unsupported, and ideas are not developed.
OR Not enough material to evaluate.

NOTE: Appropriate is defined as addressing all aspects of a topic, for example, both advantages and disadvantages, or all characteristics in questions involving choices. Furthermore, all parts of the prompt should be touched on.

RHETORICAL CONTROL

9-10
a. Introduction and conclusion effectively fulfill their separate purposes: The introduction effectively orients the reader to the topic and the conclusion not only reinforces the thesis but provides new insight.
b. Paragraphs are separate, yet cohesive, logical units. Sentences form a well-connected series of ideas or logical steps with clarity and efficiency.

7-8
a. The introduction presents the controlling idea, gives the reader the necessary background information, and orients the reader, although there may be some lack of originality in the presentation. The conclusion restates the controlling idea and provides a valid interpretation but may not provide new insight.
b. Paragraphs are usually logically developed and cohesive. Sentences are usually well-connected.

5-6
a. Introduction presents the controlling ideas but may do so mechanically or may not orient the reader to the topic effectively. The conclusion does not give the reader new insights or may contain some extraneous information.
b. Paragraphs are sometimes incompletely or illogically developed. Sentences may not be well-connected.

3-4
a. Introduction and conclusion do not restate the controlling idea. Introduction fails to orient the reader adequately, and the conclusion may not be tied to the rest of the essay.
b. Paragraphs are often incompletely or illogically developed and sentences are not well-connected.

1-2
a. Introduction and conclusion are missing or unrelated to rest of the essay.
b. There is no attempt to divide the essay into conceptual paragraphs, or the paragraphs are unrelated and the progression of ideas is very difficult to follow.

OR Not enough material to evaluate.

LANGUAGE (Grammar, Vocabulary, Register, Mechanics)

9-10
a. Except for rare minor errors (esp. articles), the grammar is native-like.
b. There is an effective balance of simple and complex sentence patterns with coordination and subordination.
c. Excellent, near-native academic vocabulary and register. Few problems with word choice.

7-8
a. Minor errors in articles, verb agreement, word form, verb form (tense, aspect) and no incomplete sentences. Meaning is never obscured and there is a clear grasp of English sentence structure.
b. There is usually a good balance of simple and complex sentences both appropriately constructed.
c. Generally, there is appropriate use of academic vocabulary and register with some errors in word choice OR writing is fluent and native-like but lacks appropriate academic register and sophisticated vocabulary.

5-6
a. Errors in article use and verb agreement and several errors in verb form and/or word form. May be some incomplete sentences. Errors almost never obscure meaning.
b. Either too many simple sentences or complex ones that are too long to process.
c. May be frequent problems with word choice; vocabulary is inaccurate or imprecise. Register lacks proper levels of sophistication.

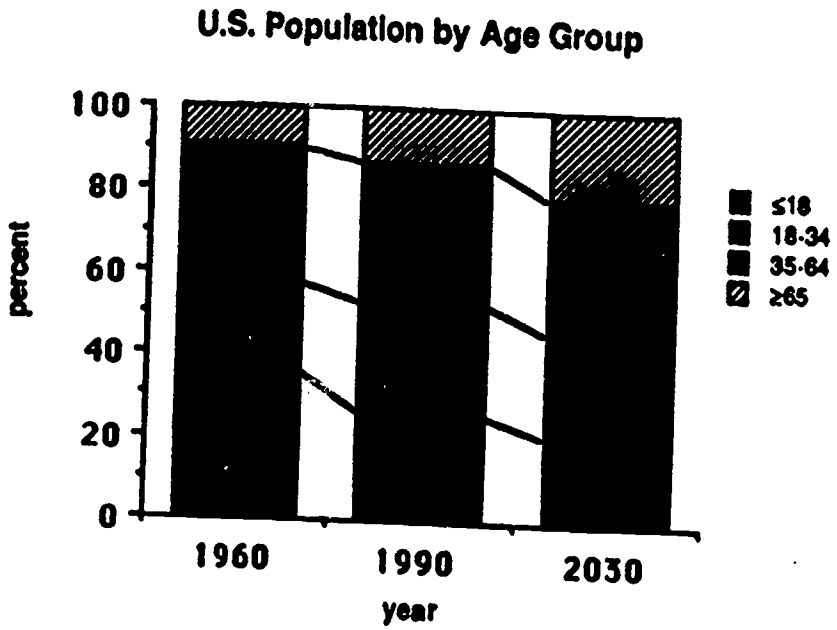
3-4
a. Several errors in all areas of grammar which often interfere with communication, although there is knowledge of basic sentence structure.
b. No variation in sentence structure.
c. Frequent errors in word choice (i.e. wrong word, not simply vague or informal word). Register is inappropriate for academic writing.

1-2
a. There are problems not only with verb formation, articles, and incomplete sentences, but sentence construction is so poor that sentences are often incomprehensible.
b. Sentences that are comprehensible are extremely simple constructions.
c. Vocabulary too simple to express meaning and/or severe errors in word choice.

OR Not enough material to evaluate.

APPENDIX B

Composition Topic 1 (GRAPH INTERPRETATION)



The above graph shows the percentage of people in different age groups in the United States population from 1960 to 2030. What does the graph tell you about changes in the population of the United States? What problems will people face as a result of these changes, and how can they best prepare for these problems?

Composition Topic 2 (CHOICE JUSTIFICATION)

Imagine that you have been offered three jobs, and you must decide which offer to accept, based on the information below. Which job would you choose, and why? Discuss the advantages and disadvantages of your choice.

	Job A	Job B	Job C
Salary	$25,000	$40,000	$100,000
Hours/week	40	50	60
Vacation	6 weeks	3 weeks	1 week
Job Satisfaction	High	Medium	Medium